yang_seonhyeHW12

```
library(data.table)
library(readr)
```

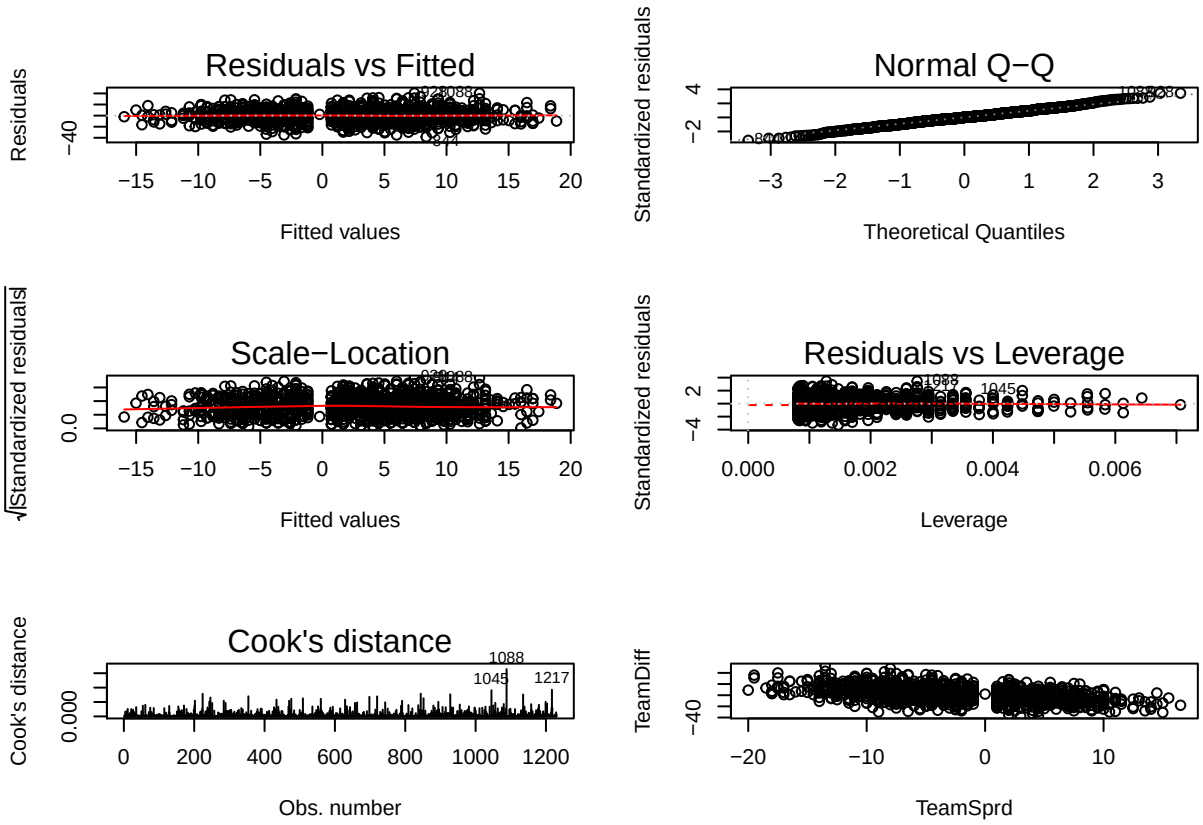
library(readr)

Question 1

```
nbaodds201415 <- read_csv("nbaodds201415.csv")
```

```
## Parsed with column specification:
## cols(
##   Datenum = col_double(),
##   Team = col_character(),
##   Dateslash = col_character(),
##   OppTeam = col_character(),
##   Home = col_double(),
##   TeamPts = col_double(),
##   OppPts = col_double(),
##   OT = col_double(),
##   TeamWin = col_double(),
##   TeamCov = col_double(),
##   TeamSprd = col_double(),
##   OvrUndr = col_double(),
##   OUCov = col_double(),
##   Team_id = col_double(),
##   OppTeam_id = col_double(),
##   TeamDiff = col_double(),
##   TotalPts = col_double()
## )
```

```
attach(nbaodds201415, warn.conflicts = F)
model<- lm(TeamDiff~TeamSprd)
par(mfrow=c(3,2))
plot(model,which=1)
plot(model,which=2)
plot(model,which=3)
plot(model,which=5)
plot(model,which=4)
plot(TeamSprd, TeamDiff)
```



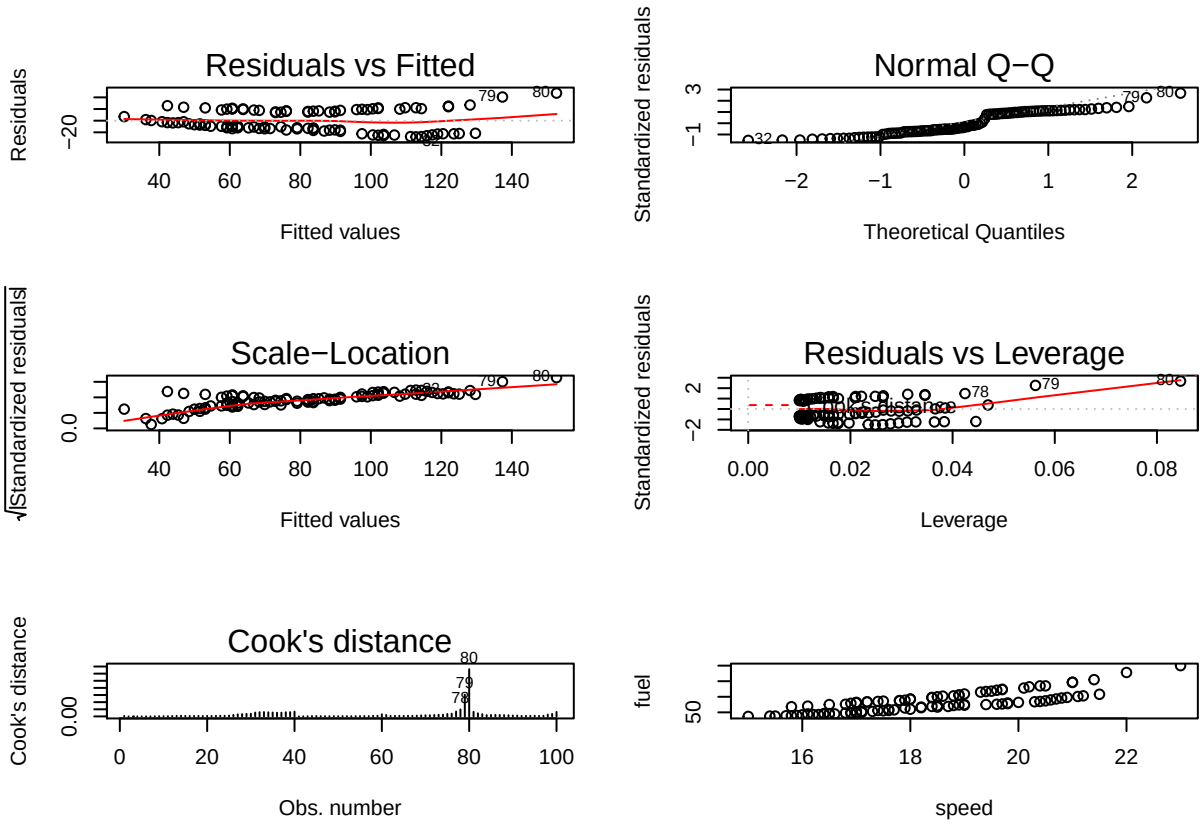
Intrepretation Our Residuals-Fitted plot shows a flat trend with roughly equal vertical spread, indicating our model has no problems. Looking at the Normal Q-Q plot, we can see a linear trend, which indicates that there is no problems with normality in this data. Moving to the Scale-Location plot, we can see that the trend is roughly flat, indicating that variance also has no problems. Finally our Residual-Leverage plot along with the plot of Cooks Distance shows that our data has few outlers and those outliers do not have much leverage, as the trendline is flat.

Question 2

```
ship_speed_fuel <- read_csv("ship_speed_fuel.csv")
```

```
## Parsed with column specification:
## cols(
##   ship_leg = col_double(),
##   speed = col_double(),
##   fuel = col_double()
## )
```

```
attach(ship_speed_fuel, warn.conflicts = F)
model1 <- lm(fuel~speed)
par(mfrow=c(3,2))
plot(model1,which=1)
plot(model1,which=2)
plot(model1,which=3)
plot(model1,which=5)
plot(model1,which=4)
plot(speed, fuel)
```



Intrepretation

For this model, our Residuals-Fitted shows a generally flat trend, but does tick upwards near the end, showing the final few points of data mightnot show a linear trend. The Normal Q-Q graph also shows a general linear trend, but points 32, 79, and 80, in addition to some values in the center are concerning, indicating potential outliers that could be eliminated. The Scale-Location graph has mostly linear trend, but has a light tick near 60, but nothing so severe as to throw out this model. The Residuals-Leverage and Cooks distance graph clearly shows our outlier data points. We can see points 78,79, and 80 sit far from our data and have significant influence (leverage) against our data, indicating that our model can be improved if we remoe them.

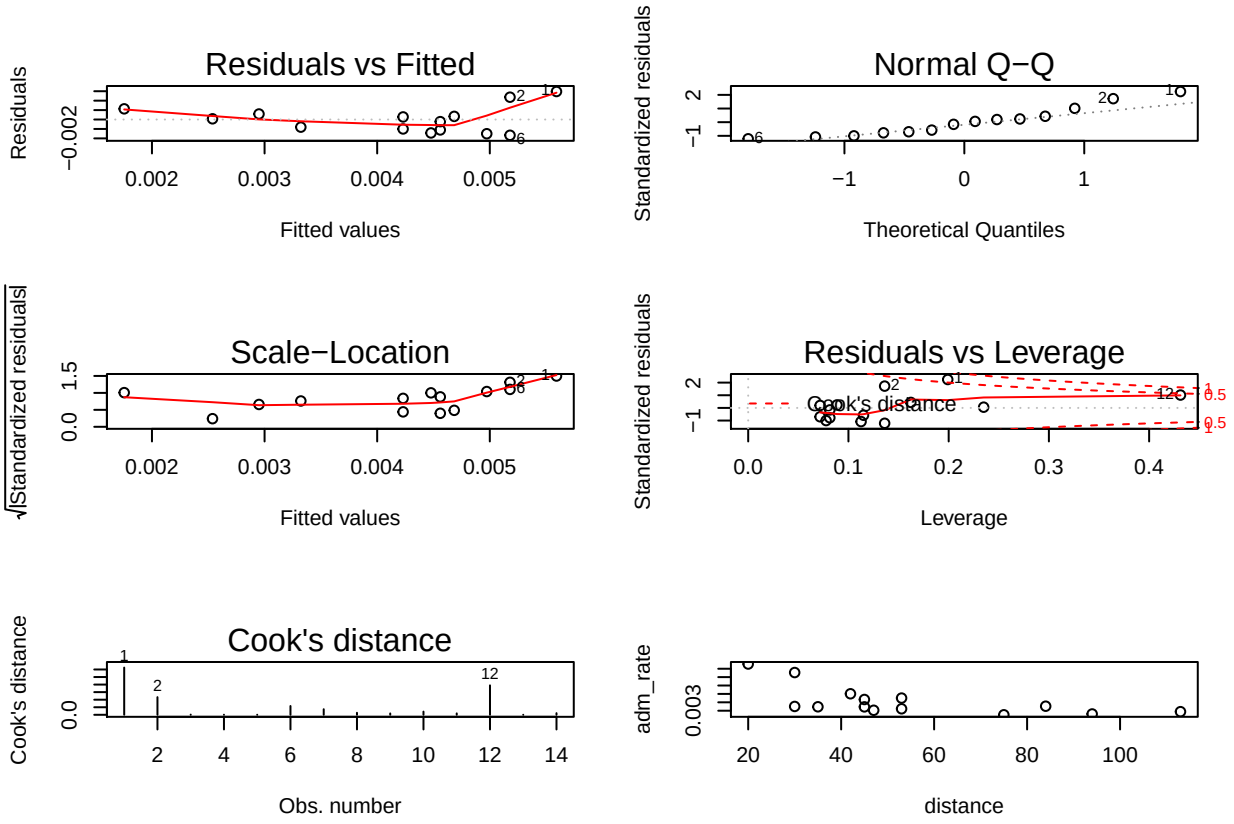
Question 3

```
asylumdistance <- read_csv("asylumdistance.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   county = col_character(),
##   pop1840 = col_double(),
##   patients = col_double(),
##   distance = col_double(),
##   adm_rate = col_double()
## )
```

```r
attach(asylumdistance, warn.conflicts = FALSE)
model2 <- lm(adm_rate~distance)
par(mfrow=c(3,2))
plot(model2,which=1)
plot(model2,which=2)
plot(model2,which=3)
plot(model2,which=5)
plot(model2,which=4)
plot(distance, adm_rate)
```



Intrepretation

From out Residuals-Fitted from Admission Rate vs Distance, we can see some concerning results. Our trendline is not flat, showing that perhaps our data is not linear at all. In addition, vertical spread is increasing as we move further down, again showing that the assumption that this model may be linar is incorrect. The Normal Q-Q plot is mostly linear, showing that there are no problems with normality. However, concerns return when looking at both Scale-Location and Residuals Vs. Leverage. Scale-Location shows no linear trend, s our residuals are not spead equally along our model, indicating a bad fit, and with our Leverage plot,we can see that each point is quite influencial, perhaps because there are so few data points. This is also reflected in the cooks distance, which shows that we have 3 very influencial points, relative to the distance of other points, which is nearly a third of our data. All of this indicates that our model may need more data points to be more effective.

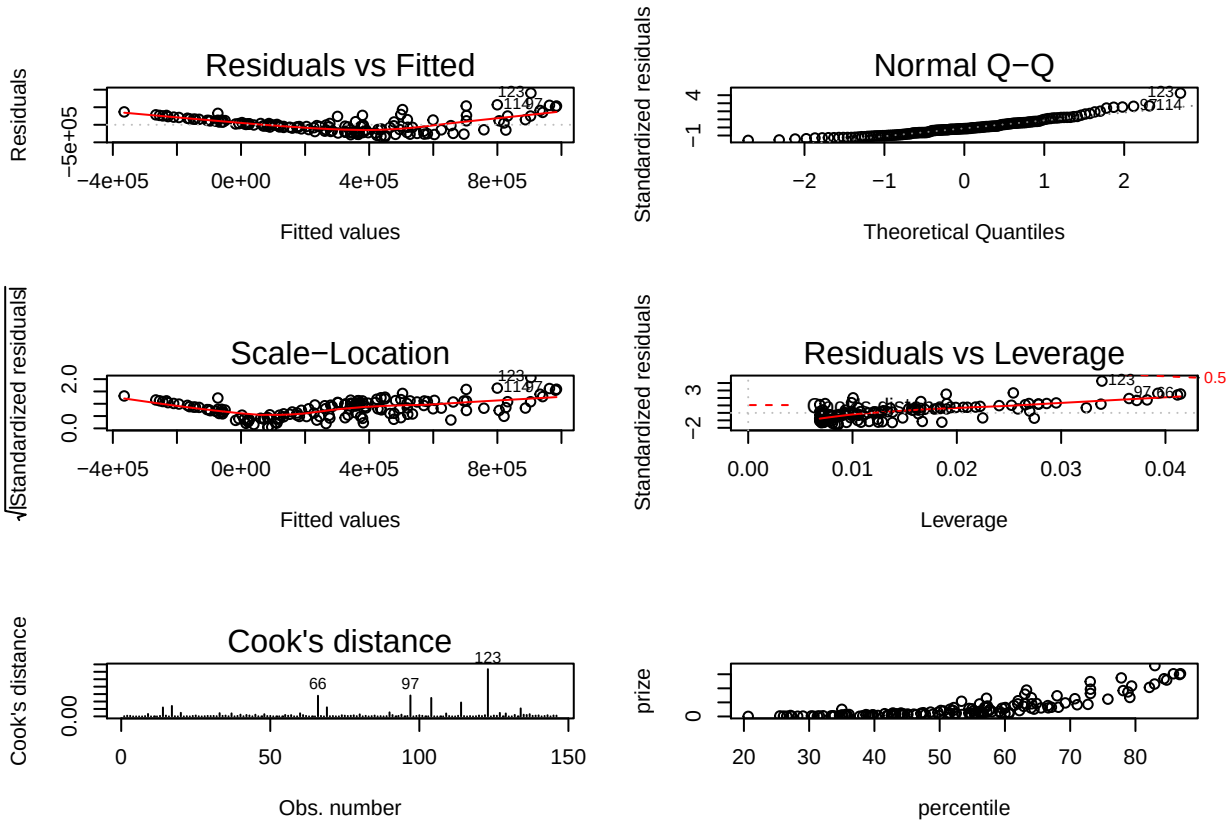
Question 4

```
lpga2009 <- read_csv("lpga2009.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   Golfer = col_double(),
##   drive = col_double(),
##   fairways = col_double(),
##   pct_greens = col_double(),
##   ave_putts = col_double(),
##   per_sandsaves = col_double(),
##   prize = col_double(),
##   `ln(prize)` = col_double(),
##   ntournaments = col_double(),
##   regputts = col_double(),
##   completed_tournaments = col_double(),
##   percentile = col_double(),
##   rounds_completed = col_double(),
##   strokes = col_double()
## )
```

```
attach(lpga2009, warn.conflicts = FALSE)
model3 <- lm(prize~percentile)
par(mfrow=c(3,2))
plot(model3,which=1)
plot(model3,which=2)
plot(model3,which=3)
plot(model3,which=5)
plot(model3,which=4)
plot(percentile, prize)
```

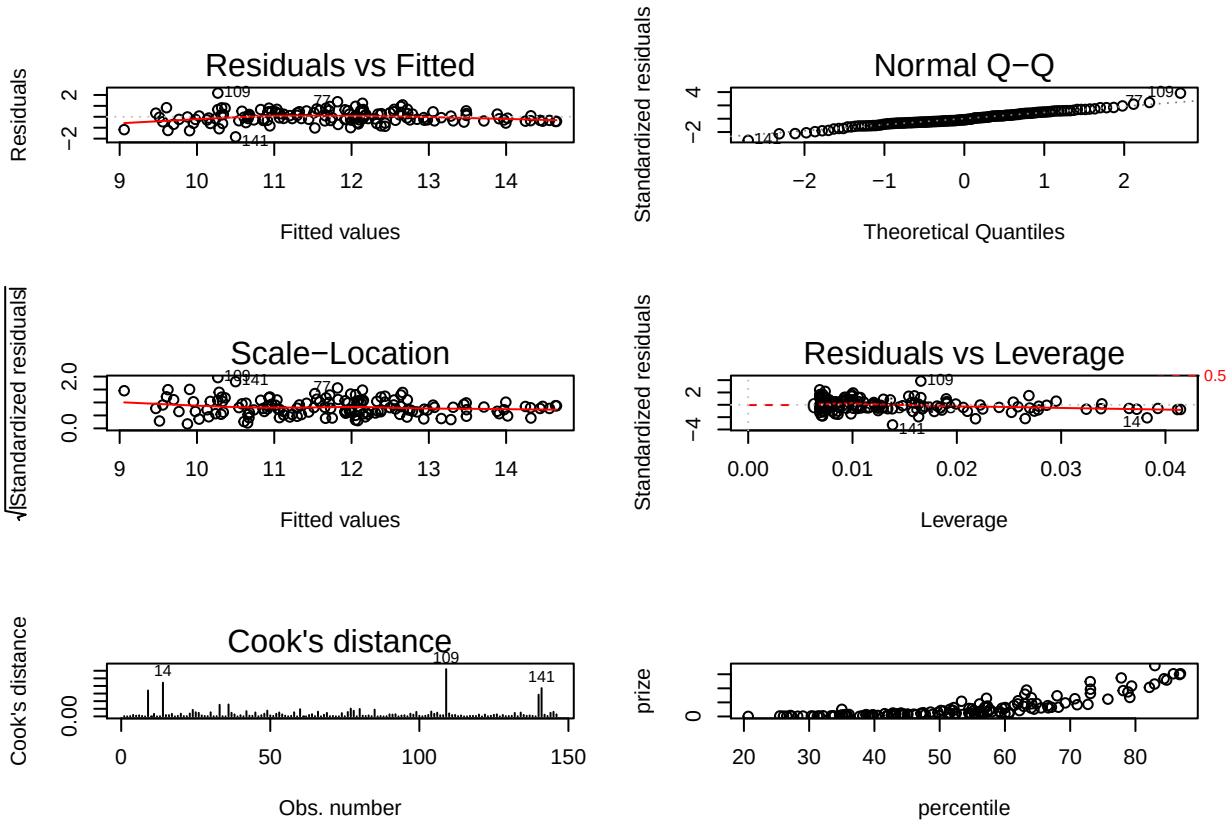


Intrepretation

Looking at the Residuals VS Fitted, we can see that there is very much a non-linear relationship here, breaking the linear assumption of our model. Our Normal Q-Q shows that the model has no normality poins. The trendline for Scale-Location is not linear, showing that there is a high degree of variation in the variance, breaking the assumption of homoscedasticity in our model. The Cooks distance and the Residuals Vs Leverage show that the top few points - people with high percentile - greatly skew our model, and have huge influnece. Because of this, we cannot properly use this model, and should either transform the data or remove high percentile players.

Question 5

```
model4 <- lm(`ln(prize)`~percentile)
par(mfrow=c(3,2))
plot(model4,which=1)
plot(model4,which=2)
plot(model4,which=3)
plot(model4,which=5)
plot(model4,which=4)
plot(percentile, prize)
```



This model simply applied a natural log transformation on our dependent variable, and that has done wonders for the predictions that we can make. Our Residuals VS Fitted graphs is far more linear and has more equal vertical spread, showing that this model is far more linear than the last. Our Normal Q-Q has remained the same, but our Scale-Location has a linear trend now. This shows no problems with variance in our model. The Residuals Vs Leverage plot is also trending linear and has fewer far leveradging points, which is further confirmed by cooks distance displayed. This then shows that a transformation has made our model more linear and able to fit more of the assumptions made.